Method	RGB	Full mIOU	Invisible mIOU
Amodal-VAE	✗	**94.68**	**62.85**
Deocclusion	✓	94.04	57.19
ResNet-Amodal-VAE	✓	94.53	61.97

Table 1: **Amodal Completion on KINS**. Checkmark on RGB column means the RGB image is used as input to the model.

Method	Full mIOU	Inv. mIOU
GT Instance Mask	87.03	0
Pred. Mask	80.83	0
Amodal-VAE	**86.03**	**60.71**
Deocclusion	84.94	52.93

Table 2: **Amodal Segmentation on KINS** Models take predicted segmentation mask as input.

Method	Full mIOU	Invisible mIOU
Amodal-VAE	94.68	62.85
Search by visible area	94.71	62.98
Search by amodal GT	95.82	69.96

Table 3: **Results for multiple predictions using samples.** We sample masks from approx. posteriors and search for best samples via visible area or full amodal GT area IOU. "Amodal-VAE" uses only the posterior mode.

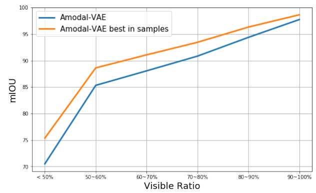

Figure 4: **Amodal Segmentation mIOU with different occlusion ratio on KINS.** X-axis represents visible area ratio. Better in zoomed-in view.

We would like to thank the reviewers for their constructive feedback.

R1&R2: Novelty. Although methodologically related approaches have appeared in the literature before, we focus on amodal object mask completion where probabilistic methods have not been explored. For this task, we are the first to propose a generative framework, which naturally captures the ambiguity inherent in the mask completion. Since naively applying a VAE is not possible, we propose a novel training strategy. Compared to previous work that uses VAEs for 3D shape completion [30], we further perform occlusion simulation and bounding box prediction, add latent code regularization during training. These contributions allow us to (1) achieve state-of-the-art results on the amodal mask completion task without amodal ground truth supervision. (2) Our mask completions outperform the drawing skills of human annotators. (3) We demonstrate the usefulness of our model on the downstream task of scene editing. (4) We qualitatively and quantitatively (Tab. 3, Fig. 4 above) show the value of explicitly modeling the uncertainty in the task.

R2&R4: Will RGB input make the task trivial? We slightly modify the model architecture and provide additional results. After the full-mask-only training stage, we use a ResNet-50 pretrained on image-net and taking RGB images as input. We concatenate the ResNet's features and the mask encoder, add two convolution layers, and predict latent code posterior distributions. As shown in Table 1, the additional RGB-based image features do not boost performance. Hence, the mask together with its object class contains enough information and we discard RGB input for simplicity.

R2: Motivation of amodal completion & Amodal instance segmentation is more interesting. We agree that amodal instance segmentation is also an interesting problem. However, amodal instance segmentation can be decomposed into instance segmentation and amodal completion. We provide additional experiments on amodal segmentation (results in Tab. 2) and will update the paper. We crop images by GT bounding box and train a ResNet-PSP segmentation model. We predict the instance masks on the test set, yielding 80.83% amodal mIOU. We use the predicted instance masks as input to perform amodal completion using both our model and the baseline (Deocclusion [37]), which we outperform. We also find that our model is robust to instance mask corruptions. Even though full mIOU drops by 8.65% (94.68% vs. 86.03%), mIOU on invisible area only drops by just 2.14% (62.85% vs. 60.71%).

R2: Lack of quantitative results for multiple posterior predictions. We agree and quantitatively evaluate this aspect (paper will be updated). For each instance, we sample 20 latent codes from the approx. posterior distribution. We calculate mIOU using masks with the best visible area IOU or best amodal GT IOU. Results show that by sampling we find masks that significantly better match amodal GT than using the approx. posterior mode (Tab. 3 above). Hence, the approx. posteriors incorporate diverse plausible masks, correctly capturing the ambiguity. This suggests the use of multiple posterior samples in downstream applications. Sampling is more beneficial for cases with significant occlusion (Fig. 4 above) and we also show (Fig. 3 of supp. material) that more occlusion correctly results in wider posteriors.

R1&R3&R4: Results are only shown on KINS dataset. We primarily focus on driving scenes and exploit two datasets with multiple classes, KINS and Cityscapes. KINS contains 7 categories (pedestrian, cyclist, person sitting, car, van, tram, truck) and Cityscapes contains 8 categories (bicycle, bus, person, train, truck, motorcycle, car, rider). We show quantitative results only on KINS, because it is the only available large-scale amodal dataset with accurate human annotations. We plan to show quantitive results on more classes and non-rigid objects for the camera-ready version.

R1 & R4: Comparisons for downstream tasks. We appreciate the suggestions. However, our main contribution is on the amodal completion task. Comparisons to other downstream task methods are beyond the scope of our paper. Nevertheless, we compared FID scores with Deocclusion [37] on the image editing task.

R1: Need to compare with more baselines & Only one category. We apologize for the confusion and will clarify this aspect in the camera-ready version. However, all experiments are conducted on datasets with multiple classes. We use Deocclusion [37] as baseline, since it is the current state-of-the-art model on amodal completion with no amodal ground truth supervision. Also, our user study demonstrates that we outperform the annotation skills of humans, which further strengthens our conclusions. We will add additional details about the user study in the final version of the paper.

R1: Transformation network. The spatial transformation network is end-to-end differentiable and trained as part of the pipeline in stages (2) and (3), using the mask reconstruction loss as supervision (we will clarify in the final version).

R3: Accuracy versus occlusion percentage . We agree. We show results in Fig. 4

Details, etc: We appreciate the suggestions and will incorporate them in the camera-ready version.